Computer Science Literature and the World Wide Web

Abby A. Goodrum, Katherine W. McCain, Steve Lawrence, C. Lee Giles
[1] College of Information Science & Technology, Drexel University
[2] NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
{abby.goodrum,kate.mccain}@cis.drexel.edu, {lawrence,giles}@research.nj.nec.com

Abstract

We analyze the computer science literature on the web and compare it to the literature indexed in the Science Citation Index (SCI). The web contains articles from throughout the research timeline, from technical reports and conference papers to journal articles and book chapters, whereas SCI focuses on journal articles. Analyzing the citation patterns of the articles, we find that journal articles and books dominate the most cited items from papers on the web and papers in SCI. However, we find that conference papers and technical reports play a very important role in computer science research, especially regarding access to the very latest research. Analysis of citations over time suggests that conference and technical report citations tend to be replaced with journal and book citations when they become available.

The web is changing the way that researchers access scientific literature. For computer science in particular, research papers are often made available on the homepages of authors and institutions. In this paper, we analyze publication and citation patterns for computer science papers on the web, and compare our results with similar analysis of computer science literature in the Science Citation Index [2, 3].

The Science Citation Index ® (SCI), created by Dr. Eugene Garfield and the Institute for Scientific Information (ISI) (`www.isinet.com`), is an index of the significant scientific journals. The SCI is created with manual assistance from human indexers, and it is expensive to index all of the literature. ISI has chosen to restrict indexing primarily to the most significant journals. The SCI began print publication in 1961 and covers about 3,500 source journals. We restricted analysis to primarily computer science literature by selecting the relevant subdivisions (Hardware & Architecture; Information Systems; Software, Graphics & Programming; and others). We accessed SCI via Dialog. In order to work around limits in the amount of data that Dialog would permit us to sort, we partitioned the data into groups of 2,000 computer science related source articles, and retrieved a systematic sample of 15 groups for further processing. Our analysis covers 30,000 source articles published between 1973 and 1999, containing about 400,000 citations [4].

ResearchIndex (`researchindex.org`), also known as CiteSeer, provides similar functionality to the SCI, in addition to other features, for literature on the web. The ResearchIndex software may be used on any database of scientific literature, however the service at `researchindex.org` currently indexes literature freely available on the publicly indexable web [6]. ResearchIndex uses Autonomous Citation Indexing (ACI) [7, 5] to create a citation index without any manual assistance, and allows researchers to perform literature search and evaluation on a database of over 300,000 computer science articles. It also provides a unique opportunity to analyze the computer science literature, much of which has not previously been available in traditional indexing services. At the time of this analysis, ResearchIndex consisted of about 200,000 articles containing about 3 million citations.

Computer science articles on the web

Figure 1 shows the distribution of articles in SCI and ResearchIndex. For ResearchIndex, the distribution is approximated from manual coding of 500 randomly selected articles, of which only 36% were cited within

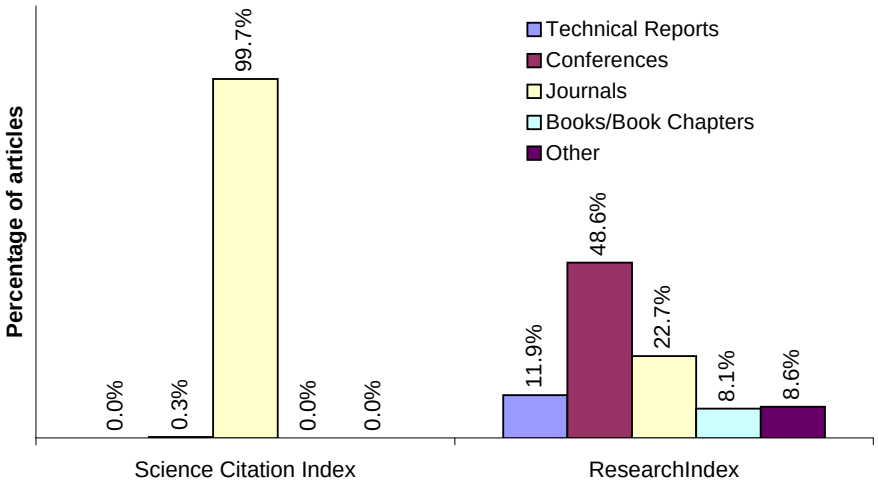

Figure 1. The distribution of article types in the Science Citation Index and ResearchIndex. The Science Citation Index focuses almost exclusively on journal articles, while ResearchIndex has a significant percentage of conference articles, books or book chapters, and technical reports. ResearchIndex represents the computer science literature available on the publicly indexable web.

the database, therefore providing easily accessible publication details (publication details for many of the remaining articles may be found in other databases, elsewhere on the web, or by contacting the authors). In addition to journal articles, ResearchIndex contains a substantial percentage of conference papers, technical reports, and books or book chapters. ResearchIndex contains articles from throughout the R&D timeline [1], whereas SCI focuses on journal articles. Figure 2 shows this graphically.

Most cited items

Figure 3 shows the distribution of the top 500 cited items in SCI and ResearchIndex. The most cited articles in both databases are books/book chapters or journal articles. ResearchIndex shows a greater percentage of conference articles amongst the most highly cited items, however conference articles still only account for 15.8% of the top 500 most cited items.

Table 1 shows the top five most cited items in each database. All but one of them are books. Of the top 25 cited items in each database, eleven occur in both lists. For a more extensive examination of cited items, see [4].

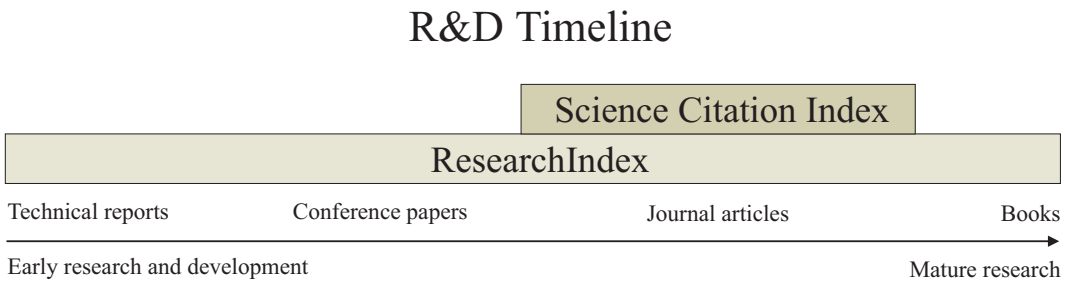

Figure 2. Research timeline. The publicly indexable web (as represented by ResearchIndex) contains articles from all points in the research timeline, complementing the better coverage of journal articles by the Science Citation Index.

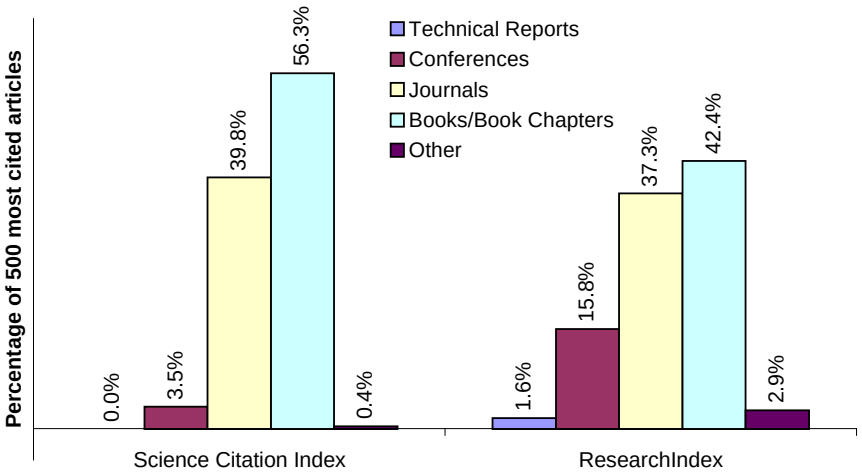

Figure 3. The distribution of article types for the top 500 most cited computer science items in Science Citation Index and ResearchIndex. Although citations to conference papers are more common in ResearchIndex, the most cited items in both databases are books/book chapters or journal articles.

Science Citation Index		**ResearchIndex**	
Citations	Article	Citations	Article
300	M. Garey, D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman, 1979.	2109	M. Garey, D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman, 1979.
254	D. Knuth. *The Art of Computer Programming*, Volume 3. Addison-Wesley, 1973.	1139	D. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.
220	A. Aho, J. Hopcroft, J. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, 1974.	1116	C. Hoare.*Communicating Sequential Processes*. Prentice Hall, 1985.
206	L. Zadeh. *Fuzzy Sets*. Information and Control, Volume 8, pp. 338–353, 1965.	1018	G. Golub, F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.
164	R. Duda, P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.	1011	T. Cormen, C. Leiserson, R. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

Table 1. The five most cited items in Science Citation Index and ResearchIndex. All but one are books.

Types of articles cited

Figure 4 shows the distribution of article types for random citations taken from articles in ResearchIndex. This shows a very different picture when compared to the distribution of most cited article types. About 34% of randomly chosen citations are for conference papers or technical reports.

Figure 5 shows the distribution of cited article types over time. Specifically, the figure plots the distribution of item types for random citations to items published a specified number of years prior to the citing publication, from zero to 20 years old. 100 random citations were classified for each age plotted. Most very recent citations are to technical reports and conference papers, while most cited items that are 10–20 years old at the time of citation are journal articles or books. For very recent journal citations, we noticed that most of them were marked “to appear” or “in press”, and did not contain specific publication details. These results suggest that conference papers and technical reports are very important to computer science research, however they tend to be replaced by citations to journal articles and books over time. Note that this replacement is often likely to not be a one-to-one mapping from an older conference paper to a more recent

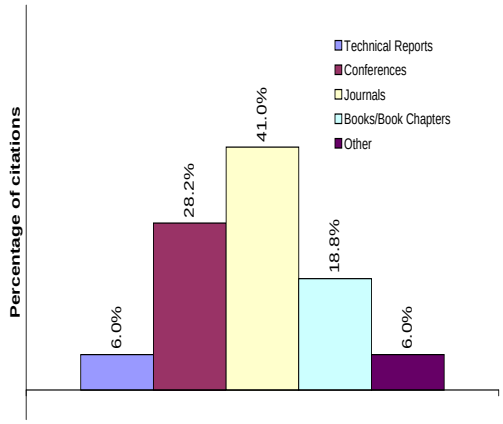

Figure 4. The distribution of random citations from articles in ResearchIndex. Conferences papers and technical reports account for a significant percentage of citations.

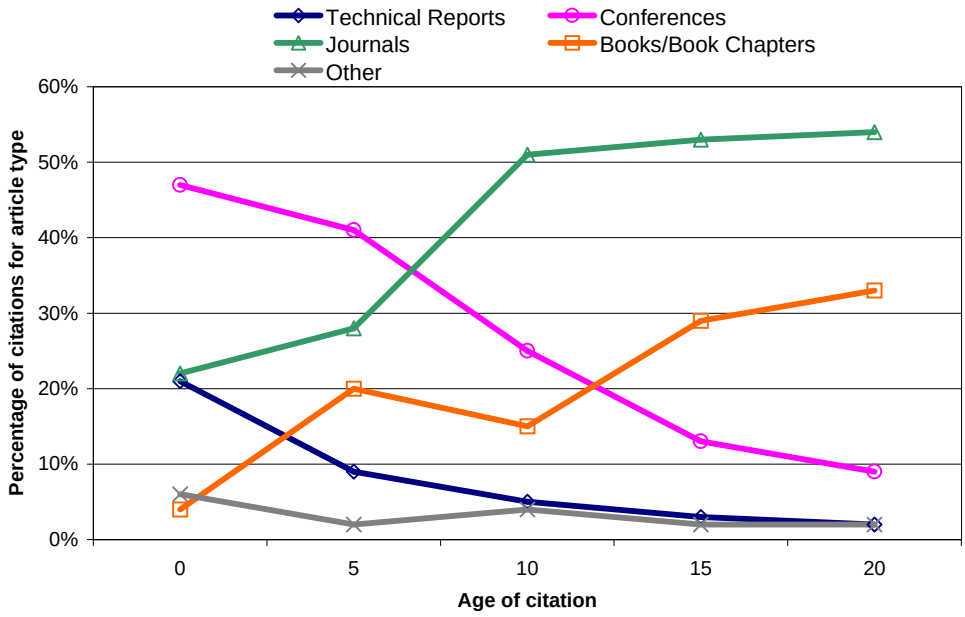

Figure 5. The distribution of article types for random citations to articles published a specified number of years prior to the citing publication. Very recent citations are predominantly to technical reports and conference articles, while most cited items that are 10–20 years old are journal articles or books.

journal paper or book. For example, journal articles and books tend to be longer and more comprehensive than conference papers, and a single journal or book citation may replace multiple conference citations. Our analysis also indicates that preprints for upcoming journal articles are very important, suggesting that it would be very beneficial for journals to make papers available online prior to paper publication. Earlier availability of publication details may also help to reduce the number of citations that do not contain these details.

Summary

Scientific communication is increasingly taking place on the web. We analyzed the computer science literature on the web and compared it to the traditional literature as indexed by the Science Citation Index.

Currently there are many articles available in each database that are not available in the other database – services like ResearchIndex complement traditional services like the Science Citation Index. Conference papers play a very important role in computer science research, especially regarding access to the very latest research. Citations to very recent articles are dominated by citations to conference papers and technical reports. However, journal articles and books dominate the most cited articles in both databases. Analysis of citations over time suggests that journal and book citations are preferred. Conference and technical report citations tend to be replaced with journal and book citations when they become available.

References

[1] S. Crawford, J. Hurd, and A. Weller. *From print to electronic: the transformation of scientific communication*. Learned Information, Medford, NJ, 1996.

[2] Eugene Garfield. Citation indexing for studying science. *Nature*, 227:669, 1970.

[3] Eugene Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Wiley, New York, 1979. ISBN 089495024X.

[4] Abby A. Goodrum, Katherine W. McCain, Steve Lawrence, and C. Lee Giles. Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management*, 2001. to appear.

[5] Steve Lawrence, Kurt Bollacker, and C. Lee Giles. Indexing and retrieval of scientific literature. In *Eighth International Conference on Information and Knowledge Management, CIKM 99*, pages 139–146, Kansas City, Missouri, November 1999.

[6] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.

[7] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.